

cassandra

ACTIVIDAD 1



**Universidad
Internacional
de Valencia**

Big Data y Data Science
Sistema de almacenamiento y gestión Big Data

Grupo: 1_A
Profesor: Pablo Suarez
Créditos: 3 ECTS
Código: 02MBID
Integrante:

1) Ronald David Jiménez Paute

Actividad 1: Creación del esquema de una base de datos orientada a columnas

Descripción:

Se pide realizar el esquema de una base de datos orientada columnas que satisfaga una serie de consultas que se desean realizar sobre el modelo conceptual. La base de datos que usaremos como referencia será Apache Cassandra. Se pide una tabla para cada consulta en la que se debe especificar que columnas serán clave primaria (partition key y clustering key). Las columnas deberán ser nombradas con el convenio siguiente: NombreEntidad_NombreAtributo. Es decir, si queremos crear una columna asociada a un atributo "Precio" de una entidad "Producto", la columna resultante sería "Producto_Precio".

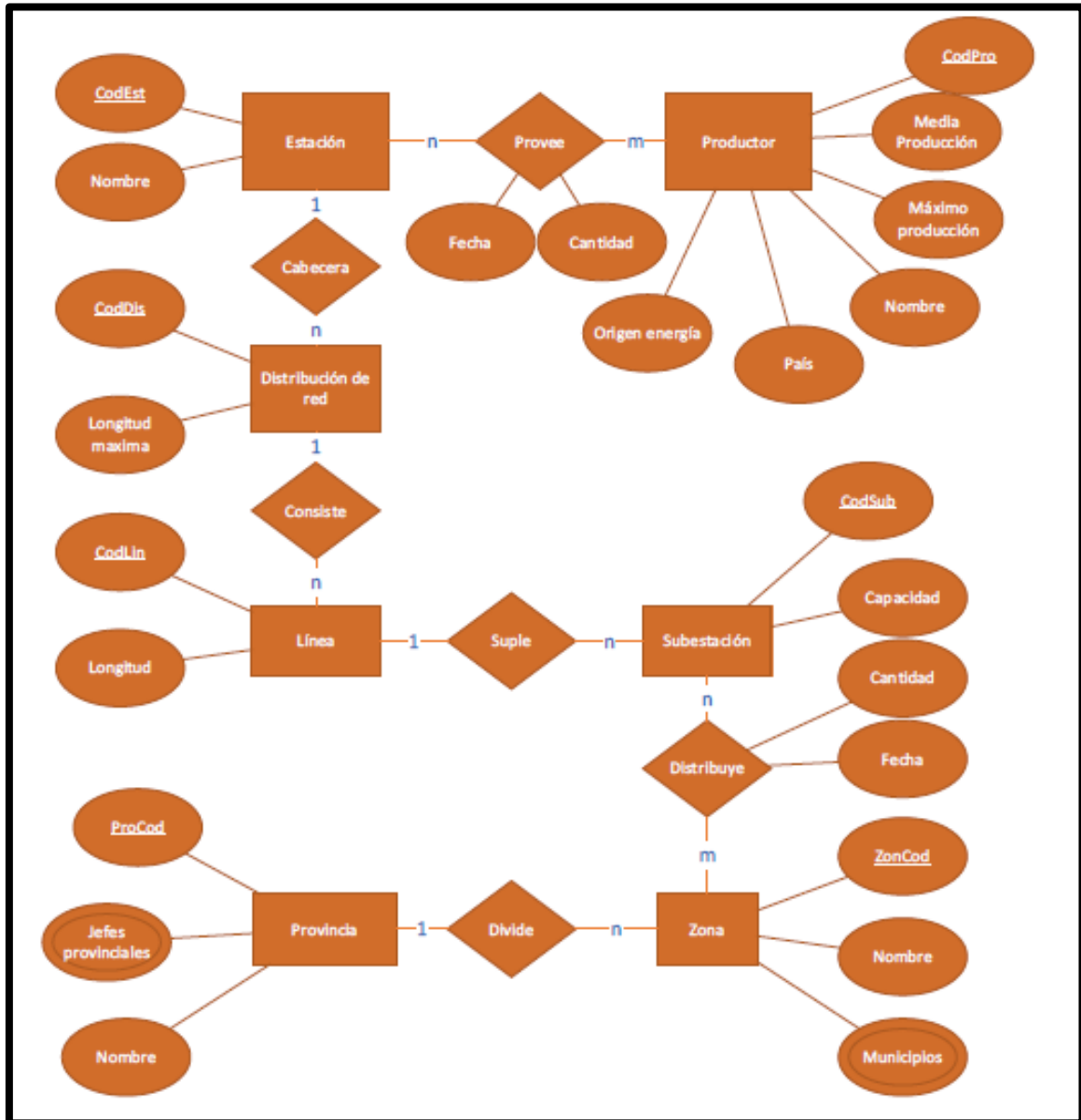
Enunciados:

1. Obtener los jefes provinciales que están asociados a una provincia buscando a través del nombre de la provincia
2. Obtener según la capacidad de cada subestación la longitud de las líneas a las que está asociado.
3. Consultar la información de la provincia a la que esté asignado un solo jefe provincial.
4. Consultar las estaciones y productores en las que estos últimos hayan provisto de una cantidad concreta en una fecha determinada.
5. Consultar según el CodDis de una distribución de red las subestaciones que esta suministra, incluyendo la longitud de la línea que supe a la subestación.
6. Obtener la capacidad sumada de todas las subestaciones que se encuentran en una zona determinada.

7. Obtener todos los productores que proveen energía buscando por la longitud de la línea a la que provee energía.

8. Buscar productores según el origen de la energía que producen (eólica, nuclear, carbón, solar o gas). Nota: En la actualidad el 50% de productores usan carbón, un 30% nuclear, un 15% solar y un 5% eólica.

Modelo Conceptual:



Desarrollo:

1. Obtener los jefes provinciales que están asociados a una provincia buscando a través del nombre de la provincia.

Column	Primary Key
Provincia_Nombres	PK
Provincia_ProCod	CK
Provincia_Jefes	

Justificación: En este enunciado no existe relación, solo hay una entidad para la consulta. Se podrá considerar que la columna "Provincia_ProCod" cumple la función de Clustering Key, es cierto que los valores son únicos. La columna "Provincia_Nombre" se podría hacer la búsqueda de datos, siendo así cumple la función de Partition Key y son valores que se repiten.

2. Obtener según la capacidad de cada subestación la longitud de las líneas a las que está asociado.

Column	Primary Key
Subestacion_Capacidad	PK
Subestacion_CodSub	CK
Linea_CodLin	
Linea_Longitud	

Justificación: Según en los casos especiales se menciona que cuando existe la relación 1:n, requiere incluir la Partition Key de la entidad que posee n. Respecto del modelo conceptual, en el atributo "Subestacion_capacidad" de la entidad Subestación, cumple la función de los casos especiales. Por consiguiente, se podría considerar como Partition Key. La columna "Subestacion_CodSub" se consideró como Clustering Key para ayudar con la unicidad de los datos. El resto de la columna se consideró como atributo normal.

3. Consultar la información de la provincia a la que esté asignado un solo jefe provincial.

Column	Primary Key
Provincia_JefeProvincial	PK
Provincia_Procod	CK
Provincia_Nombre	
Provincia_Jefes	

Justificación: En este caso no existe relación, solo hay una entidad para la consulta. Respecto del enunciado se podría considerar que la columna "Provincia_ProCod" cumple la función de Clustering Key, es cierto que los valores son únicos. La columna "Provincia_JefesProvinciales" es un atributo conjunto pero lo ideal sería denominar como atributo normal (Provincia_JefeProvincial). Por consiguiente, se podría considerar como Partition Key. El resto de la columna se consideró como atributo normal.

- Consultar las estaciones y productores en las que estos últimos hayan provisto de una cantidad concreta en una fecha determinada.

Column	Primary Key
Est_Prod_Fecha	PK
Est_Prod_Cantidad	PK
Estacion_CodEst	CK
Productor_CodProd	CK
Estacion_Nombre	
Productor_OrigenEnergía	

Justificación: Según en los casos especiales se menciona que cuando existe una relación n:m, requiere incluir la Partition Key de ambas entidades relacionadas. Respecto del modelo conceptual, para que haya unicidad de los datos se podría considerar Clustering key de ambas entidades. Por consiguiente, se consideró la columna de "Estacion_CodEst" y "Productor_CodProd" como Clustering Key. La columna de "Est_Prod_Fecha" y "Est_Prod_Cantidad" como Partition Key. El resto de la columna como atributo normal.

- Consultar según el CodDis de una distribución de red las subestaciones que esta suministra, incluyendo la longitud de la línea que supe a la subestación.

Column	Primary Key
Distribucion_CodDis	PK
Subestacion_CodSub	CK

Distribucion_LongitudMáxima Linea_Longitud Linea_Codlin Subestacion_Capacidad	
--	--

Justificación: Según en el caso de modelado se menciona que si existen tres entidades A, B y C todas relacionadas con relaciones 1:n encadenadas, solo se necesita la clave primaria de C para garantizar unicidad. Respecto del modelo conceptual, hay tres entidades relacionadas con una relación 1:n (Distribución, Línea, Subestación). Por consiguiente, la columna "Distribucion_CodDis" se consideró como Clustering Key para ayudar con la unicidad de los datos. La columna "Linea_Longitud" se consideró como Partition Key para hacer la búsqueda de datos. El resto de la columna se consideró como atributo normal.

- Obtener la capacidad sumada de todas las subestaciones que se encuentran en una zona determinada.

Column	Primary Key
Zona_ZonCod NumCapacidad	PK +

Justificación: Según en los casos especiales se menciona que cuando sirve para almacenar información de con cuantas instancias de una entidad, se debe agregar con el signo +. Respecto del enunciado, se agregó otra columna "NumCapacidad" con el signo + para acumular información. Para la búsqueda de datos, la columna "Zona_ZonCod" se consideró como Partition Key,

- Obtener todos los productores que proveen energía buscando por la longitud de la línea a la que provee energía.

Column	Primary Key
Linea_Longitud Productor_CodPro Distribucion_CodDis Linea_CodLin	PK CK CK CK

Estacion_CodEst Productor_Nombre	
-------------------------------------	--

Justificación: Según de los casos especiales se menciona que si existen dos o más entidades todas relacionadas con relaciones diferentes, solo se necesita Clave Primaria de cada entidad o de relación que poseen n:m. La columna "Linea_Longitud" se consideró como Partition Key para hacer la búsqueda de los datos. La columna de "Productor_CodPro", "Distribucion_CodDis", "Linea_CodLin" se consideraron como Clustering Key. El resto de la columna como atributo normal.

8. Buscar productores según el origen de la energía que producen (eólica, nuclear, carbón, solar o gas). Nota: En la actualidad el 50% de productores usan carbón, un 30% nuclear, un 15% solar y un 5% eólica.

Column	Primary Key
Productor_OrigenEnergía	PK
Productor_CodPro	CK
Productor_MáximoProducción	CK
Productor_Nombre	
Productor_MediaProducción	
Productor_País	

Justificación: Respecto del modelo conceptual no existe la relación, solo especifica una sola entidad. Para la búsqueda de los datos, la columna "Productor_OrigenEnergía" se consideró como Partition Key. Para que haya unicidad de los datos, la columna "Productor_CodPro" se consideró como Clustering Key. Para obtener el porcentaje según el origen de la energía, se estableció "Productor_MaximoProducción" como Clustering Key. El resto de la columna se estableció como atributo normal.

Referencias:

Transparencias de clases expositivas

Manual de la asignatura "Sistemas de almacenamiento y Gestión Big Data"